IMLS Digital Collections and Content

Grant LG-02-02-0281

Interim Performance Report 2
1 April 2003 – 30 September 2003

Submitted by Timothy W. Cole, Principal Investigator, November 2003
University of Illinois at Urbana Champaign

University of Illinois
1301 W. Springfield
Urbana, IL 61801
Tel 217.244.7809
Fax 217.244.7764

Grant LG-02-02-0281
Interim Performance Report 2
1 Apr. 2003 – 30 Sept. 2003

Submitted by Timothy W. Cole, Principal Investigator, Nov. 2003

Summary

The IMLS Digital Collections and Content project has made good progress in all areas of the project during the previous six months. As noted in prior report, a few early milestones were delayed due to longer than anticipated time to obtain clearance for survey instruments, but surveys were distributed to 92 National Leadership Grant (NLG) projects in mid-September. We anticipate completing all milestones, albeit on a delayed schedule. In August, a collection level metadata schema was approved by the Steering Committee and submitted to IMLS for approval. Progress was made on design of the collection registry interface and database structure. The project team worked with several institutions to either set up or advise on implementation of Open Archives Initiative (OAI) metadata provider services. As of Oct. 23, 2003, the item level metadata repository holds 43,462 records from fifteen NLG projects. Repository is currently searchable through an early alpha interface. Project research team has begun interviewing participants from selected projects, and continues to host a bi-weekly metadata roundtable.

General Project Activities

Financial report

The Annual Financial Status Report is attached. In addition to grant award expenditures shown on the Report, the UIUC Library has contributed 10% of the annual salary and fringe benefits for project PI, Timothy Cole, 5% of the annual salaries and fringe benefits for co-PI's Nuala Koetter and William Mischo, and 3% of the annual salary and fringe benefits for consultant, Beth Sandore. The Library also contributed funding for travel to the Open Forum on Metadata Registries in Santa Fe, NM in Jan. 2003 and the ALA Conference in Toronto, Canada in June 2003.

The first year allocation of the grant award has not been fully spent due to several factors including delays (2 - 4 months) in hiring and starting project staff members. Unanticipated delays were also encountered in obtaining formal clearance for project survey instruments. Subject to further schedule adjustments, we anticipate that we will request a 3 or 6 month no-cost extension to compensate for these delays. In addition, the co-PIs leading primary research activities, Carole Palmer and Michael Twidale, claimed only half of the salary budgeted for their contributions to the project in year 1. In addition the research team delayed their computer purchase to project year 2. We expect to spend the excess salary funds from the RA and research faculty lines on an additional research assistant for Jan. – Aug. 2004.

Timeline

The original timeline for the project has been substantially changed due to the delay in distributing the survey instrument. Included in Appendix One is a new schedule of completion. New dates for milestones are noted throughout this Interim Performance Report.

Dissemination

The IMLS DCC project has been presented in a number of forums. Copies of these can be found in Appendix Two.

Sarah Shreeves presented "Integrating Resources for Information Discovery" at the Digital Resources for Cultural Heritage: Current Status, Future Needs. A Strategic Assessment Workshop in Washington, D.C. Aug. 25, 2003.
http://imlsdcc.grainger.uiuc.edu/Shreeves_IntegratingResources.ppt

Tim Cole presented "IMLS NLG Collection Registry & Item Level Metadata Repository at the University of Illinois" and "Notes on Panel on Future of OAI" at the 4th Open Archives Forum Workshop in Bath, UK on Sept. 4, 2003.
http://dli.grainger.uiuc.edu/Publications/TWCole/OAForumWkshpBath/ColeOAFWkshpBath2003.ppt
http://dli.grainger.uiuc.edu/Publications/TWCole/oaforumwkshpbath/ColeOAPanelBath2003.ppt

The IMLS DCC project had two posters at the 2004 Dublin Core (DC) Conference in Seattle, WA on Sept. 29-Oct. 2, 2003: "Tracking Metadata Use for Digital Collections" by Ellen Knutson, Carole Palmer, and Mike Twidale and "Developing a Collection Registry for IMLS NLG Digital Collections" by Sarah Shreeves and Tim Cole.
http://www.siderean.com/dc2003/705_Poster43.pdf
http://www.siderean.com/dc2003/706_Poster49-color.pdf

In Mar. 2004 we are planning to hold a workshop on the Open Archives Initiative at Webwise in Chicago, IL as well as present on our work thus far. We are actively considering other near-term dissemination venues including the Joint Conference on Digital Libraries 2004 and the American Society for Information and Technology Annual Conference 2004.

Steering committee activity

The Steering Committee met via conference call on July 29, 2003 to discuss the proposed collection description metadata schema. The Steering Committee website has been updated regularly with relevant documents produced by the IMLS DCC project.[1] The next meeting of the Steering Committee is tentatively planned for Mar. 5th, 2004 after the Webwise Conference in Chicago, IL.

Collection Registry Metadata Schema and Service

Survey of IMLS NLG projects

In early Sept., the Office of Management and Budget approved our survey of and plan for follow-up emails with relevant IMLS NLG projects. On Sept. 15th a packet with two survey instruments - the first collecting and verifying project and collection information (for initial registry entries) and the second supporting our research investigations - were sent to the principal investigators (PIs) of 92 NLG projects. We created a SQL database to record the results of the surveys as they are returned.

[1] See http://imlsdcc.grainger.uiuc.edu/steeringcommittee/ (password protected).

During Oct. and Nov. 2003, we will contact non-respondents first by email and then by phone and will continue to enter survey data into our database. The survey results will allow us to create preliminary records for the collection registry, categorize NLG projects according to their viability for implementing OAI data provider services, and provide information for our research.

Developing the collection-description metadata schema

Much of the project's work in the last six months was concentrated on further developing the collection-description metadata schema. Our work was informed by our participation in ongoing discussions on the Dublin Core Collection Description Working Group listserv[2] and meeting at DC-2003 and by the conversations at the Metadata Roundtable, a bi-weekly meeting of faculty and students interested in metadata issues held at the Graduate School of Library and Information Science. We captured much of our process in creating the schema in our poster and poster abstract for the DC-2003 conference. In June 2003 three NLG projects tested the IMLS DCC collection description metadata schema. There were no unexpected findings from this test run; the participants seemed to understand the schema. After examination and discussion of the CIDOC Conceptual Reference Model[3] (a top-level ontology and proposed ISO standard for the semantic integration of cultural information) we made further adjustments to the schema including elements identifying the physical collection(s) from which the digital collection(s) was derived.

In July 2003 we presented the revised schema to the IMLS DCC Steering Committee for their approval. We convened a meeting via conference call of steering committee members on July 29, 2003 to vet the schema. The Steering Committee agreed in general that the July 2003 revision of the collection description schema was appropriate and was not missing any major elements. Some minor changes, deletions, and additions were made. A final copy of the metadata schema was circulated to the Steering Committee in mid-Aug. for final review and then was submitted to IMLS for approval on Aug. 20, 2003. Documentation of this process can be found in Appendix Three.

Designing and building the collection registry

A preliminary version of the database for the collection registry was built in Aug. 2003 and was tested using the collection descriptions submitted by our testers. We developed a preliminary 'staff' interface to aid in navigating the relationships between collections, projects. We shared this interface with IMLS on Sept. 3, 2003. In addition we began designing a public interface to the collection (again based on the three test records). We examined other collection registries such as Cornucopia (http://www.cornucopia.org.uk) and Enrich UK (http://www.enrichuk.org) for functionality and interface design features. We built a browse screen based on the GEM subject headings, a short display, and a full display.[4] This alpha mock-up was shared with IMLS on Sept. 10, 2003.

During the next six month period we will build preliminary collection description records in the collection registry using the survey results. We will design, develop, and test the forms that will enable NLG projects to enter and maintain collection metadata and, pending OMB approval, ask NLG projects to verify and augment their collection description records using these Web forms. Our revised estimate for a beta version of the collection registry (pending approval by OMB) is Mar. 2004 for WebWise in Chicago. An initial production version of the registry should be available by June 2004.

[2] See the listserv archives: http://www.jiscmail.ac.uk/lists/DC-COLLECTIONS.html.
[3] See http://cidoc.ics.forth.gr/.
[4] See http://imlsdcc.grainger.uiuc.edu/collections/Gemtop.asp (password protected). Also included in Appendix Three.

Item-Level Metadata Repository

Assisting projects in implementing OAI-data provider services

Although we were handicapped by the delay in distributing the survey, we did continue discussions with several NLG projects about implementing OAI data provider services, and collaborated during the last six months with two more projects to make their metadata harvestable.

Static OAI data provider service

In July 2003 we set up an OAI static repository for the NLG project "American Natural Science in the First Half of the Nineteenth Century" based at the Academy of Natural Science. A recent development in the OAI protocol and designed for use with small, relatively static metadata collections, a static OAI repository is a single XML file which contains metadata records and which sits on the data provider's standard web server. A third party acts as a gateway through which an OAI service provider can then harvest that static XML file the metadata. This obviates the need for the source data provider to implement a new dynamic web service. A full technical description of the static gateway can be found at http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm. The project team worked with Eileen Mathias to map metadata from MARC records to simple DC and produced a single XML file (with both MARC and DC records available for harvest) which is now available through our third party gateway.[5] This success of this implementation indicates that the static provider service is a good solution for institutions lacking technical infrastructure to implement new, dynamic web services.

ContentDM OAI-data provider service

In July 2003 we worked with the Washington State Libraries to harvest metadata from their ContentDM data provider service. ContentDM is a digital library management system which has built in an OAI data provider service. However, the current version of ContentDM (3.5) does not support resumption tokens. These are an optional feature in the 2.0 OAI protocol but aid in 'flow control' by allowing a data provider to issue records in manageable chunks to a service provider, thus limiting the peak load on both systems. Although optional, the implementation of resumption tokens is particularly important for large data providers. We examined other possible avenues for harvesting these records. We determined that dividing metadata into smaller sets (maximum of 10,000 records per set) could facilitate harvesting without flow control. We also developed a successful workaround in which we harvested records individually. While this work-around was slow, it put little to no stress on the web server and all metadata records were harvested successfully. We have contacted ContentDM about the lack of full functionality in their turn-key OAI data provider service. (In addition to not implementing resumption tokens, ContentDM can only provide metadata in simple Dublin Core).

Other OAI provider implementation discussions

In addition to the Academy of Natural Science and the Washington State Libraries, we consulted with several other NLG grantees including the Missouri Botanical Gardens, University of Connecticut, Indiana University, University of Washington, and Illinois State Library regarding plans for setting up OAI data provider services.

We also are tracking why NLG projects might not be able or ready to implement data provider services. Survey results will help with this task. A preliminary review based on conversations held so

[5]OAI base URL:
http://imlsdcc.grainger.uiuc.edu/gateway/oai.asp/www.acnatsci.org/library/collections/imls/nlg/AcadNatSciStatic.xml.

far indicate that NLG projects may not be in a position to implement OAI data provider services because:

- There is no item level metadata. This is true for many exhibit and learning object focused projects.
- The collection is not yet public. NLG projects wish to wait until they unveil their digital collection before sharing the metadata.
- Infrastructure is not in place. The metadata may not be mapped into Dublin Core or stored in such a way to set up OAI data provider services.
- The technical infrastructure is in transition or will be in transition. NLG projects are reluctant to implement OAI provider services in the midst of a migration to a new content management system.
- Agreement must be reached among all project collaborators to share metadata via OAI.

During the next six months we will use survey results to segment NLG projects into four groups. Our preliminary results (as of Oct 24, 2003) indicate the following breakdown:

- **Group 1** - Projects with OAI data provider sites for NLG content: 15.
- **Group 2** - Projects whose institutions have an OAI implementation (not yet being used for NLG content) and NLG projects that have explicitly expressed plans to add OAI functionality: 16.
- **Group 3** - Projects who meet certain technical criteria - e.g. have item-level metadata and a maintained web site: 6.
- **Group 4** - Projects with no item-level metadata or no interest in providing metadata via OAI: 6.
- **Unknown:** 50

Metadata harvesting and design of item-level repository

We have continued to harvest metadata from OAI-compliant NLG projects into our alpha item-level repository. As of Oct. 23, 2003, we have harvested approximately 43,462 DC records from 15 OAI-compliant NLG projects.[6] The repository is available at http://imlsdcc.grainger.uiuc.edu/searchimls/ (password protected). We have made some slight adjustments to the interface of this repository and have shared it with the NLG projects we are harvesting.

Over the next six months, we will continue to harvest OAI-compliant NLG projects. We plan further enhancements to the interface of the item-level repository, including the ability to search in specific subject areas (using the GEM subject headings). We also plan on using Spotfire, a data analysis tool, to aid us in analyzing metadata harvested and identifying areas where we may need to normalize values.

Research

Data collection

The research plan for year one consisted of four iterative stages of data collection and analysis: 1) content analysis of the NLG proposals, 2) Survey 1, 3) e-mail follow-up survey, and 4) phone interviews with a representative group of projects. Stage 1 was completed as expected, but due to the delay in survey distribution we decided to make accommodations in the original plan. This involved moving forward with stage 4 without the benefit of having the baseline of data from the survey. Initially, interviews were to follow preliminary analysis of the survey and follow-up results, to build on and enrich the necessarily brief responses provided by survey methods. To keep the project moving forward, we altered our research design and in recent months began gathering interview data from

[6] See http://imlsdcc.grainger.uiuc.edu/steeringcommittee/NLGprojectsharvested.doc for projects and number of records harvested. (password protected). Also see Appendix Four.

some NLG projects. However, because surveys were sent at the same time we began interviews, several questions were asked in both instruments and therefore the iterative, longitudinal questioning approach was not achieved.

As of Oct. 16, 2003, 13 interviews, conducted with participants from 9 project sites, have been completed. Transcription of the interviews is well underway, and we have begun initial analysis, especially in the area of collection definition and application of metadata schemes. These results were presented at the Dublin Core 2003 conference. See below for more details.

As mentioned above, the survey results are beginning to come in and are being entered into a SQL database. As of the writing of this report the response rate was at slightly more than twenty percent. Non-respondents have been contacted by the project coordinator which should greatly increase the response rate. Once the survey data has been reviewed, we will be sending email follow-up questions to clarify and expand on the survey questions.

Our next major data collection activity will be conducting focus groups in Mar. 2004 at WebWise in Chicago. We plan to conduct two focus groups of approximately 6-10 participants each. The participants will be a convenience sample of IMLS NLG grantees in attendance at WebWise. We have begun the OMB approval process, and a final package of focus group questions and details on the research method will be sent to IMLS by Dec. 1, 2003.

Dissemination of research results

As mentioned above, the research team gave a poster presentation at the Dublin Core conference in Seattle, WA, Sept. 29 – Oct. 2, entitled “Tracking Metadata Use for Digital Collection.” We reported on Stage 1 results from the content analysis of the project proposals and preliminary results from the interviews. It was evident from comments made in Neil McLean's plenary session and from the dearth of user-based projects, that there is a great need for research of this type in the metadata community. A copy of the poster is included in Appendix Two.

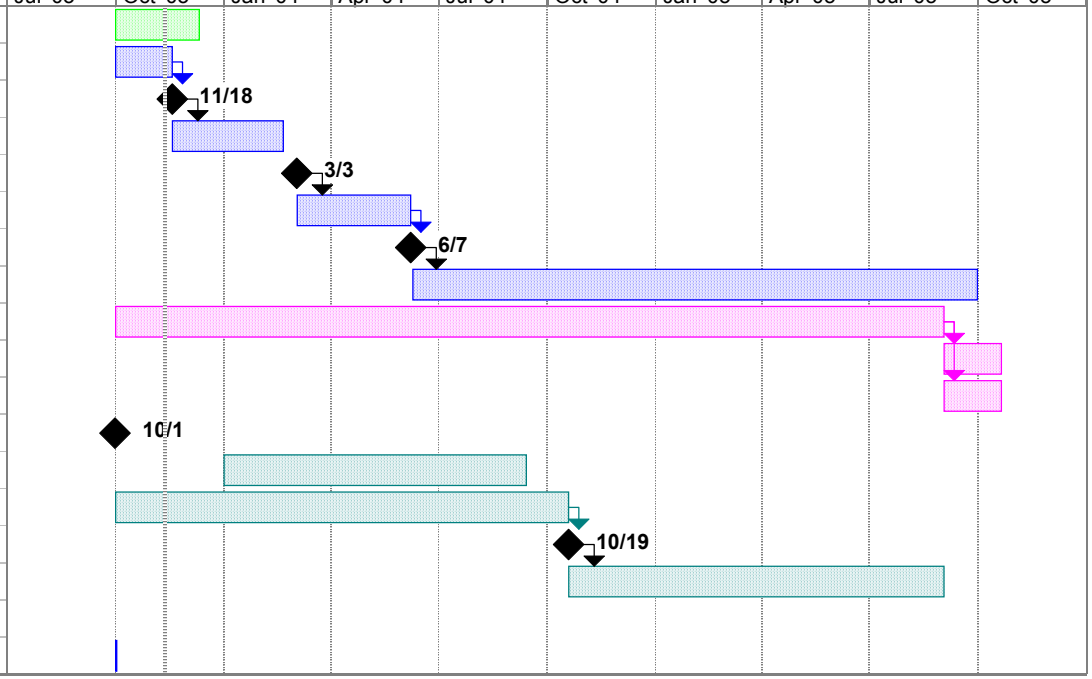
We expect to present further results at conferences in the coming year in the form of contributed papers. Possible venues include ASIS&T Annual Meeting and the 2004 JCDL Conference. The team is also in the beginning stages of writing a paper on the topic of collection definition for digital distributed repositories.

Related Activities

We have been conducting a bi-weekly metadata roundtable where members of the Graduate School and Library and Information Science and the University Library community meet to discuss issues that surround the use and creation of metadata. Some of the topics we have discussed include collection level metadata, the Dublin Core recommended values for the collection type property, the Dublin Core Collection Level Application Profile, CIDOC Conceptual Reference Model (CRM), and the definition of a collection. Regular participants include faculty and both masters and doctoral students from GSLIS as well as university librarians. Guest participants have included Jane Greenberg from the School of Information and Library Science, University of North Carolina at Chapel Hill, who is a member of our steering committee, and other GSLIS visiting scholars. Both Carole Palmer and Ellen Knutson attended the Collection Description Working Group at the Dublin Core conference in Seattle. The concepts discussed and ideas generated at both the Collection Description Working Group Meeting and at the metadata roundtables have informed the research and implementation processes of the project and the research team's plans for publication on the topic of collection definition.

Appendix One - IMLS DCC Schedule of Completion

ID	Task Name
1	Initial Survey of IMLS Grantees
2	Define Collection Level Metadata Schema
3	Template for Registry Information *
4	Implement Collection Registry
5	Online Access to Collection Registry (BETA)
6	Testing and Refinement of Registry Service
7	Online Access to Collection Registry (Initial PRODUCTION) *
8	Continued Testing and Refinement of Registry Service
9	Assist with Provider Implementations
10	Report on Barriers to Interoperability *
11	Report and Recommendations on Repository Marketing Plan *
12	Metadata Repository Search Available(Beta)
13	Analyze available metadata
14	Testing and Refinement of Repository
15	Metadata Repository Search Available (Initial Production)*
16	Continued Testing and Refinement of Repository
17	
18	* Indicates a deliverable specified in the Cooperative Agreement



Appendix Two: Dissemination Activities

Posters:

Knutson, E., Palmer, C. & Twidale, M. (2003). Tracking Metadata Use for Digital Collections [Poster Abstract]. In *DC-2003: Proceedings of the International DCMI Metadata Conference and Workshop p. 243-244.*

⇒ Abstract
⇒ Poster

Shreeves, S.L. & Cole, T.W. (2003). Developing a Collection Registry for IMLS NLG Digital Collections [Poster Abstract]. In *DC-2003: Proceedings of the International DCMI Metadata Conference and Workshop p. 241-242*.

⇒ Abstract
⇒ Poster

Presentations:

Shreeves, S.L. "Integrating Resources for Information Discovery". Digital Resources for Cultural Heritage: Current Status, Future Needs. A Strategic Assessment Workshop. Washington, D.C. August 25 2003

Cole, T.W. "IMLS NLG Collection Registry & Item Level Metadata Repository at the University of Illinois" and "Notes on Panel on Future of OAI". 4th Open Archives Forum Workshop. Bath, UK. September 4 2003.

Appendix Four – National Leadership Grant Collections and Number of Records Harvested

Academy of Natural Sciences
"American Natural Science in the First Half of the Nineteenth Century" - LL-90013
347 records

Colorado Digitization Program
"Heritage Colorado" - LL-90094
18,824 records

Florida Center for Library Automation
"Florida Environmental Information Online" (Part of "Linking Florida's Natural Heritage" - LL-80016)
1,155 records

Tufts University
"Bolles Archive of London" - ND-00015
35 records

University of Georgia
"Southeastern Native American Documents" - LL-90019 and ND-00017
164 records

University of Illinois
"Teaching with Digital Content" - NL-00003
1,983 records

University of Maine
"Maine Music Box" - LG-03-02-0116
1,596 records

University of Michigan
"Flora and Fauna of the Great Lakes"- NL-00034
12,988 records

University of Minnesota
"Summons to Comradeship: World War I and II Posters" - ND-10007
725 records

University of North Carolina
"Southern Homefront" - LL-80202
403 records

University of North Carolina
"North Carolina in Black and White" - ND-00031
422 records

University of Tennessee
"Tennessee Documentary History"
1,216 records

University of Wisconsin-Madison
"Africa Focus" - LL-80131
100 records

Washington State University
"Columbia River Basin Ethnic History" - NL-10032
3,504 records